Informative Value of Individual and Relational Data Compared Through Business-Oriented Community Detection

Vincent Labatut, Jean-Michel Balasque

Informative Value of Individual and Relational Data Compared Through Business-Oriented Community Detection

Vincent Labatut and Jean-Michel Balasque

Abstract Despites the great interest caused by social networks in Business Science, their analysis is rarely performed in both a global and systematic way in this field. This could be explained by the fact their practical extraction is a difficult and costly task. One may ask if equivalent information could be retrieved from less expensive, individual data (i.e. describing single individuals instead of pairs). In this work, we try to address this question through group detection. We gather both types of data from a population of students, estimate groups separately using individual and relational data, and obtain sets of clusters and communities, respectively. We measure the overlap between clusters and communities, which turns out to be relatively weak. We also define a predictive model, allowing us to identify the most discriminant attributes for the communities, and to reveal the presence of a tenuous link between the relational and individual data. Our results indicate both types of data convey considerably different information in this specific context, and can therefore be considered as complementary. To emphasize the interest of communities for Business Science, we also conduct an analysis based on hobbies and purchased brands.

1 Introduction

Bringing new insights in decision-making analysis, social networks have raised a great interest in the scientific community, including Business Science. However, in this field, they have a paradoxical position: on the one hand the interest of network analysis has been greatly emphasized years ago, but on the other hand this tool is not very widespread yet. In the first part of this section, we review previous works focusing on network analysis for Business and Marketing Sciences and try

Vincent Labatut
Galatasaray University, Computer Science Department, Çırağan Cad. No:36, 34357 Ortaköy/İstanbul, Turkey
e-mail: vlabatut@gsu.edu.tr

Jean-Michel Balasque
Galatasaray University, Business Science & Marketing Department, Çırağan Cad. No:36, 34357 Ortaköy/İstanbul, Turkey
e-mail: jmbalasque@gsu.edu.tr

to find explanations to this observation. Then, in the second part, we explore more thoroughly one of these explanations, which is related to the nature of the data requested to extract social networks, and derive the problematics and core ideas of this paper.

1.1 Network Analysis in Business Science

In Marketing Science, and more generally in all the fields of Business Science, the strength of the concept of social network can be considered at different levels [1]. First, locally, by taking into account the interaction between a person and his precise relational context, it constitutes a good tool to better understand individual decisions. Second, at the level of a whole system, it provides a meaningful analysis basis and offers the necessary information to improve both global organization and individual activities management. The main point in both research and practice has been the possible benefits a person or a firm can get from social networks. Consequently, their analysis has been elaborated primarily in a utilitarian perspective, with a particular emphasis on their impact on the nature and efficiency of information dissemination. Their role was noticeably studied in the context of competitiveness in the construction sector [2], firm innovativeness [3], investors attraction for venture capital [4], effective use of their social capital [5] in the labor market [6], and administrative board decisions [7], among others. In Marketing Science, the focus has been put more on speed of information diffusion, with a major interest in word-of-mouth [8], changes of opinions and adoption of innovations inside groups of people (mainly consumers and potential consumers) [9-11] or diffusion of specific products [12].

In most of these studies, the analysis is centered on a single or a few persons, and consists in studying their most immediate connections in great details. Even if the investigation concerns a whole social system (e.g. group or firm), the focus remains local. Some works study the role these individuals of interest have in the network. Other works analyze the possible effects of the social network on these individuals, and generalize the resulting observations to the rest of the network, or to some subgroups of persons. This approach can be criticized in several ways. First, influence processes within social networks vary considerably depending on the nature, structure and strength of the links that connect the considered persons. For instance, Steyer *et al*. [10] showed the efficiency of information dissemination processes, used for viral marketing, depends on the whole network structure. Van der Merwe [11] described its effect on the role of opinion leaders. According to various authors [13,14], both opinion spreading and speed of innovation adoption depend on the considered network structure and density. Second, the interest of adopting a non-local approach is backed by several Marketing studies like [15], which, following a stream of Sociology studies, emphasized the necessity of taking sub-networks or cliques into account. These structures diffuse information faster: people belonging to them are more quickly and more deeply influenced,

they rapidly adopt new products. So, from a managerial point of view, they are of higher interest. In the context of complex networks, this naturally leads to the notion of community, i.e. a group of nodes with denser relationships, compared to the rest of the network.

Burt [16] and Perry-Smith [17] showed structural holes improve the emergence of new ideas. In their analysis of firm innovativeness, Simon and Tellier [3] differentiated two kinds of innovativeness: exploitation and exploration. They showed groups with denser inner-connections were more efficient to diffuse ideas, but rupture innovations were less likely to happen in those parts of the firm. These denser groups can also be seen by people as stressful areas which constrain them too much, preventing any behavior opposed to the dominant one. In this context, people sometimes rely on persons not belonging to their community [18] or weakly connected [19]. Then, detecting communities can also, by contrast, reveal the zones of lower link density, or structural holes. It additionally allows a deeper study of the network structure by performing a centrality analysis. Indeed, people located in-between communities often play specific roles because of their central position [20]. Interestingly, this last point brings us back to the local approach, illustrating how complementary they are: a global approach can be used to locate persons of interest, which can then be studied more attentively.

Besides this complementary nature and the fact a global approach seems necessary to improve our understanding of social networks and their effects, it is rarely adopted in the fields of Marketing and Business Sciences. Moreover, when it is the case, authors generally do not use a systematic method. For instance, the works cited in the previous paragraphs [3,17] do not use a precise definition of the concept of community and do not intend to identify all communities present in the studied network. We see two possible reasons for this. First, this kind of analysis is computationally far more demanding than local approaches. It relies on relatively new tools (both theoretically and practically speaking), making intensive use of modern computers. Because of this novelty, they do not have penetrated Business Science deeply yet. Second, and more importantly, practical extraction of social networks is a difficult and costly task [21,22], because the information it requires is often difficult to access and thereby expensive.

1.2 Nature and Cost of Data

Let us consider data according to two axes: the cost axis and the individual vs. relational axis. In the latter, *individual* refers to data describing only one person, whereas *relational* points out data concerning two (or more) persons. On the first axis, we can distinguish three kinds of data, differing both by the nature of the information they convey and on how difficult and costly they are to obtain.

First, *factual* information is the most easily accessible; it corresponds to acknowledged, generally publicly available, facts. For individual data, we can cite

for example social status, gender, age, etc. For relational data, it can take the form of communication streams such as email exchanges, lists of collaborations, etc.

Second, what we call *behavioral* information can either result from observations or be obtained directly by interrogating the persons of interest. For individual data, it describes how some person reacts to a given situation, whereas for relational data, the concern will be put on interactions between people, for instance by measuring the time workers spend together in a firm.

Third, *sentimental* information is related to feelings and thoughts. It is the most difficult to retrieve, since it cannot be accessed in other way than more or less direct questions, or very advanced physiological techniques [23,24]. For individual data, it is for instance brands representations, firm image or products preferences. For relational data it corresponds to feelings (friendship, love, hate, admiration…) people have for each other. Sentimental relational data can be estimated through questions of the sociometric form, where each person is asked to list his acquaintances and to quantify the strength and orientation of their relationships [25]. This so-called sociometric approach is considered to be both the most efficient, in terms of quality of the retrieved relationships, and the most difficult to apply [21]. Extracting a social network requires relational data, which is globally more difficult and costly to gather than individual data [26]. Indeed, most available factual data focus on single persons (resumés, archives, surveys…); observing interactions in a whole population obviously requires more resources than concentrating on a single individual; and making people speak about others can be an even more sensitive task than making them reveal personal details.

From this data-related difficulty regarding social networks extraction, a question arises: can the information conveyed by social networks be retrieved by other, less expensive, means? In this work, we try to tackle this issue through the angle of group detection. We analyze data coming from a survey conducted on a population of university students. Its questions targeted both relational data, with a sociometric approach, and individual data, including factual, behavioral and sentimental-centered questions. However, in this article we present only the first stage of our work, which is concerned with the relational data and only a part of the individual data (mainly factual).

From the relational data, we extract a social network, in which we detect communities. We then analyze them from a Business-oriented perspective, and show their practical importance in this field of research. In parallel, we perform a classic cluster analysis on the factual individual data, in order to obtain clusters of students. The question is then to know if this analysis, which is standard in Business Science, gives access to the same information than community detection, which is much less employed. For this purpose, we first compare individual and relational information through an analysis of similarities and differences between the two kinds of groups. We then use our results from the cluster analysis to design a predictive model able to estimate the community of an instance based on its individual attributes. This allows us to identify which attributes are the most important to characterize the communities, and therefore to tackle the problem of community composition by analyzing them in terms of individual data.

Our contributions are both practical and theoretical. First, we present and analyze some original data, and interpret the results of this analysis in the context of Business Sciences. Second, the problem of comparing the informative value of individual and relational data was not raised before, to our knowledge, and we propose an original method to tackle it. The rest of this article is organized as follow. In section two, we describe the survey we set to collect data, focusing on the parts used in the present work. We also give a short description of the tools used to analyze them: community detection, cluster analysis, and our predictive modeling approach. In section three, we present and comment our results regarding the identified communities and clusters, their characteristics and usefulness. In the conclusion, we highlight the original points of our study, discuss its limitations and explain how it can be continued.

2 Methods

We conducted two different analyses. The first is a comparison of groups estimated independently from the individual and relational data, resulting in so-called clusters and communities, respectively. The second is a study of the community composition in terms of individual factual data. In this section, we describe first how we gathered the data, and which part of them was used in this study. We then present the methods used to estimate and compare the groups of students, and finally the predictive analysis approach we applied to study community composition.

2.1 Data Collection

The data used in this article are a part of some results obtained from a larger survey. In this subsection, we first present the general survey and context of the study, and we then focus on the data selected to be used in this work.

2.1.1 Survey and Context

The Galatasaray University (GSU) is a small Turkish public institution of about 2000 students, located in Istanbul, near the Bosphorus. It offers a wide variety of courses (sociology, economics, international studies, management, philosophy, computer science, engineering, law…) taught mainly in French. In Turkey, students enter universities after having passed a national competitive examination called ÖSS. The ranking they get at this occasion is very important, because it has a direct effect on the set of universities and departments they can choose to study in. The GSU is one of the top universities in several fields, and as such it attracts

students with very high rankings. For most students, the name of the university itself is more important than the actual standard of the department they are going to enter. This particular university can also recruit students directly from Turkish French-speaking high schools, thanks to a specific internal examination. Approximately two thirds of the students are undergraduates and will get a Lisans (i.e. License, or BS) diploma, the rest being Master and PhD students. Each department has a promotion of about 30 students per level. Community and cultural life is highly developed; the university counts forty active sports clubs or cultural associations. There is a very strong feeling of belonging to a group, enhanced by the fact the name Galatasaray also refers to a prestigious high school, a popular association football club, and various other cultural and sporting structures. After the university, very strong ties remain between GSU alumni, which usually help each other professionally.

In this context, we have conducted a study on the social network of current GSU students. A university can be considered as a relatively close system for students, in the sense most of their friends also belong to it, making it an appropriate field of investigation. Accordingly to the previous description, this seems to be particularly true for the GSU. Our study is based on a survey taking place at several periods, in order to be able to study some of the network dynamics. The results presented here are limited to data obtained during the first phase of the overall research project, which took place during spring 2009 and involved 224 respondents, mainly at the Lisans level.

We designed a questionnaire focusing on social and personal attributes, social interactions (especially in the daily university environment), purchasing behavior and favorite brands. The questions can be distributed into three different thematic parts, although this separation does not appear in the questionnaire, voluntarily. The first one concerns factual data: age, gender, clubs or associations membership (inside and outside the university), school situation, previous high-school. The second part focuses on the student's behavior relatively to his friends: nature of the communication means he uses (cell phone, Facebook, Skype...); and also concerning his shopping habits, information sources, buying behavior. The third part concerns his feelings about the university, his vision of his relationships with his friends, his desires, hobbies, goals and favorite brands. All questions were designed to gather individual data (i.e. information limited to the student himself), except one, which was dedicated to relational data (i.e. data involving two students). We adopted a classic sociometric approach, consisting in asking the student to name the peers he finds the most important in his everyday life, and to quantify these relationships on a scale ranging from –5 (hate) to +5 (love). A website was created to gather the responses. Part of the required information was very personal and sensitive, so a specific procedure was set to guarantee perfect anonymity, replacing all names by meaningless codes.

2.1.2 Selected Data

As stated before, in this article we focus only on a part of the gathered data. First, the relational data (sociometric question) are used to build the social network, which means it is based on the feelings each student declares to have about his fellows. Each node in the network represents a student which either responded to the survey or was cited by a respondent (and sometimes both). Consequently, the network contains more nodes (622) than we had respondents (224), since some cited persons did not answer during this phase of the survey. Each link is directed from the respondent towards the cited student, and has a weight corresponding to the score the respondent associated to the relationship. The presence of a link between two nodes represents the fact the respondent considers the cited student as one of his most important fellows. Therefore, the communities identified in this network correspond to groups of people affectively bound inside the university.

Second, factual individual data are used to estimate clusters of students, in order to be subsequently compared to communities. The complete list of factual individual attributes is given in Table 1. In the rest of the document, we will refer to these data simply as 'the attributes'.

Table 1. Factual individual attributes used for both clustering and predictive analyses

Attribute	Type	Description
Gender	Dichotomous	Male vs. Female
Department	Nominal	The GSU has 22 departments
Class	Ordinal	Current year (Preparatory and Lisans): 6 different levels
Grade	Real	Current grade of the student, expressed from 0 to 4
Entrance	Dichotomous	Entrance examination: National vs. Internal
High-School	Nominal	High-school name
Category	Nominal	High-school type: 6 different categories
City	Nominal	High-school city: 55 different cities
Specialization	Nominal	High-school specialization: 17 categories
Clubs	Dichotomous	Forty activities inside and outside the GSU

Third, three additional attributes were selected to illustrate how communities can be used in the context of Business Science. They correspond to behavioral information and concern the students' hobbies and the brands of mobile phone and digital player they own. All three attributes are nominal, and they are not used during the cluster analysis.

2.2 Analysis Tools

To identify groups (clusters and communities), we used a set of representative algorithms. We chose to apply several algorithms in order to be able to compare their results and ensure group stability. In other terms, our goal was not to compare the algorithms in terms of performance, but rather in terms of agreement, and to identify the most consensual groups. For this reason, we selected the algorithms by considering first the nature of the groups they detect and/or the process applied to perform this detection, so that a wide scope of approaches were represented. To ensure the reliability of the results, we favored proven algorithms, i.e. tools widely used in previous published works. Finally, we also chose the algorithms depending on whether their implementations were publicly available or not. All of the selected tools output a partition of the analyzed data, which means every instance belongs to exactly one group (groups are therefore mutually exclusive). Algorithmic complexity and scalability were not an issue, since the processing was performed completely off-line and on limited data.

Group detection algorithms differ mainly in the method they use for group identification. Some adopt a hierarchical approach, which can be either agglomerative or divisive. In the first case, each object is initially considered as a group, and the algorithm merges them iteratively until only a single group remains. In the second case, on the contrary, the process starts with a single group containing all objects, which is iteratively separated in subgroups until each of them contains only one object. In both case, the characteristic point is the criterion used to select the groups to be merged or divided. Other algorithms rely on the so-called partitional approach, which consists in defining an initial partition, randomly or according to some approximate method, and then iteratively improving it, relatively to some criterion, by moving objects from one group to the other. Finally, besides these general approaches, ad hoc methods exist, which we will describe in greater details in the following.

In this subsection, we first describe the cluster analysis and community detection algorithms we applied on our data. We then explain how we compared their results to assess their agreement. Finally, we formally define the model we used to predict community membership in function of the identified clusters.

2.2.1 Cluster Analysis

One of the most important points in the detection of clusters (based on the individual data) is the choice of an appropriate dissimilarity function, allowing to properly compare the instances. For some of our attributes (Gender, Class, Grade and Entrance), this choice was straightforward because their nature does not let much freedom: they are simple binary, ordinal and numeric values. But the remaining ones (Department, Clubs and Highschool-related attributes) have specificities requiring to make some methodological choices. For this reason, we de-

fined and tested several functions, and present here only the one giving the best results.

The Department attribute has a nominal value, and we should consider two departments as different if they are not represented by the same symbol. However, a department belongs to a faculty, and can be considered as thematically closer to another department from the same faculty than to a completely unrelated one. For this reason, we chose to consider two departments as partially similar if they belong to the same faculty. The Clubs information is problematic because it actually is a very sparse vector of binary values, and we could not consider them as separate attributes, or their importance would be overstated. This is why we decided to use Jaccard's coefficient [27] to summarize all club memberships under the form of a single value. The problem of the Highschool-related attributes is they are highly correlated, which is why we also had to combine them under the form of a single value. Using our expertise of the context, we defined a synthetic nominal attribute corresponding to different highschool categories, by considering the highschool city, type of education, and teaching language, and the student highschool specialization.

To our knowledge, no previous work is supporting the fact some combinations of factual attributes would have a better predictive power than others relatively to the composition of the communities. Consequently, there is no *a priori* reason to favor a certain subset of attributes to perform the cluster analysis. We therefore opted for an exploratory approach, and considered all possible combinations of factual attributes during the cluster analysis.

Table 2. Summary of the cluster analysis algorithms applied to the individual data

Algorithm	Approach	Reference	R library
Agnes	Agglomerative	[28]	cluster
DBScan	Density-based	[29]	fpc
Diana	Divisive	[28]	cluster
Pam	Partitional	[28]	cluster
TwoStep	Hybrid	[30]	homemade

The five algorithms we selected are summarized in Table 2. They are standard and proven tools, available in many data mining softwares and representative of the different families of cluster analysis methods. We used some implementations defined in two R language [31] libraries as indicated in Table 2, except for TwoStep, which we programmed ourselves.

Pam (Partitioning Around Medoids) [28] uses a partitional approach, and necessitates to know *a priori* the number of clusters. For each cluster, an instance is initially randomly selected and considered as its center. Each remaining instance is then assigned to the cluster whose center is the most similar. After this step, any instance in a cluster might become its new center if this allows reducing the sum of all instance-to-center dissimilarities. The process is repeated with these new

centers until there is no more change in clusters and centers, leading to a stable partition.

Agnes (Agglomerative Nesting) [28] follows an agglomerative hierarchical approach. First, each instance is considered as a cluster. Then an iterative process is applied, which merges the least dissimilar clusters. Assessing this dissimilarity is straightforward if both clusters contain only one instance. Otherwise, the average linkage is used, i.e. the dissimilarity between the clusters is the average of all instance-to-instance dissimilarities. The process ends when only one cluster containing all instances remains.

Diana (Divisive Analysis) [28] is also hierarchical, but unlike Agnes it is divisive. It starts with a single cluster containing all instances, which will be split in two. First, the instance with the highest average dissimilarity to all other instances is identified and considered as the seed for a new cluster. The instances of the original cluster are then considered one by one, in a way similar enough to the process implemented in Pam, so that instances more similar to the new cluster are reassigned to it. The same splitting method is then iteratively applied to the largest cluster until all clusters contain only one instance. The largest cluster refers here to the cluster with higher diameter, i.e. maximal dissimilarity over all its pairs of instances.

TwoStep [30] is a very general method consisting in applying successively two different algorithms, with at least a hierarchical agglomerative algorithm as the second one. We used Diana for the first step and Ward's method [32] for the second. The first step can be considered as a preprocessing phase, consisting in identifying the dense areas of the data space in order to produce the smallest clusters of interest. During the second phase, larger clusters are built by merging these, in order to improve the quality of the partition.

DBScan (Density-Based Spatial Clustering of Applications with Noise) [29] is a density-based algorithm, which allows it to uncover non-convex clusters. The process starts by randomly selecting an instance, and considering its neighborhood. If it is dense enough, the instance is the seed of a new cluster, in which all its neighbors are also placed. Each neighbor is then considered: if its own neighborhood is dense enough, its own neighbors are added to the cluster and the process goes on with them. Once all possible nodes have been added to the cluster, one of the remaining nodes is selected randomly to start a new cluster using the same principle.

2.2.2 Community Detection

The selected community detection algorithms are summarized in Table 3. The interested reader will find a more detailed description of their functioning in this subsection. But before, we need to introduce Newman's *Modularity* measure [33], which is used by several of them. It estimates the quality of a network partition relatively to its topology. The original formulation is based on a normalized com-

munity adjacency matrix whose elements e_{ij} represent the proportion of links between communities i and j [33]:

$$Q = \sum_i \left(e_{ii} - e_{i+} e_{+i} \right) \tag{1}$$

Where e_{i+} and e_{+i} are the sums over row and column i, respectively. The term e_{ii} corresponds to the observed proportion of links inside community i. The term $e_{i+}\ e_{+i}$ is an estimation of the same quantity for a network whose links are randomly distributed. The modularity consequently measures how much the considered network differs from a random network, in terms of number of intra-community links and relatively to the partition of interest. It can be considered as a chance-corrected measure. The theoretical maximum value is 1, but it is related to the network structure, and in practice it cannot be reached for all networks. When considering real-world networks, partitions whose modularity reaches 0.7 are considered to be very good [34,35]. Note numerous modularity variants exist, some of them allowing to process weighted [36] and directed [37] links.

Table 3. Summary of the community detection algorithms applied to the data. The + signs represent the ability to process directed (D) or weighted (W) links (note these abilities depend on both the algorithm and the considered implementation). The last column displays the type of implementation we used: iGraph R library [38] or author's implementation.

Algorithm	Approach	D	W	Reference	Implementation
CommFind	Laplacian, Spectral, Aglomerative	–	–	[39]	Author
EdgeBetweenness	Edge-betweenness, Divisive	–	–	[40]	iGraph
EigenVector	Modularity, Spectral, Divisive	+	–	[41]	iGraph
FastGreedy	Modularity, Greedy, Hierarchical	–	+	[42]	iGraph
InfoMap	Compression, Simulated annealing	+	+	[43]	Author
LabelPropagation	Information propagation	–	+	[44]	iGraph
Louvain	Modularity, Greedy, Hybrid	–	+	[45]	Author
MarkovCluster	Random-walk	+	+	[46]	Author
Radicchi	Link-transitivity, Divisive	–	–	[47]	Author
SpinGlass	Modularity, Simulated annealing	–	+	[48]	iGraph
WalkTrap	Random-walk, Agglomerative	–	+	[49]	iGraph

Some algorithms are completely based on the modularity measure, and maximize it using various means. Exhaustive optimization is computationally intractable though, due to the number of possible partitions, so they perform an approximate processing. Modularity is also used in most hierarchical algorithms to select the best cut in the generated hierarchy of partitions, and therefore determine the optimal number of communities. Moreover, in section 3 we use modularity to compare the various estimated community structures in terms of quality.

FastGreedy [42] is historically the first algorithm designed to maximize modularity. It relies on a hierarchical agglomerative approach to perform a greedy optimization. An iterative process takes place, starting with a partition containing as many communities as nodes. At each iteration, two communities are selected and merged to obtain a new partition. They are selected so that the modularity of this partition is maximal. The process ends when the merge leads to a single community containing all nodes. *Louvain* [45] also greedily optimizes modularity using a hierarchical agglomerative approach. The main difference with FastGreedy is the application of a partitional step at each iteration, allowing to merge several communities at once. This can affect the higher level communities, resulting in potentially different partitions.

EigenVector [41] is also a hierarchical algorithm, but unlike FastGreedy and Louvain, it is divisive. Moreover, it relies on a completely different optimization method inspired by spectral bisection [50]. This classic graph-partitioning approach takes advantage of some spectral properties of the Laplacian matrix, which is derived from the graph adjacency matrix. In the case of EigenVector, a so-called modularity matrix is used instead, which can be considered as an adjacency matrix undergoing the same chance-correction than the one used for the modularity measure. The algorithm then considers the eigenvector associated to the highest eigenvalue, and splits the network in two depending on the signs of its elements. This results in an approximate optimization of the modularity. This division is repeated iteratively until each community contains only one node. *CommFind* [39] also uses a spectral approach, but this one is based on the traditional Laplacian matrix. Instead of using only the best eigenvector, it selects the few best ones, which allows taking more information into account. It can then apply a cluster analysis on the resulting data, instead of iteratively performing bisections. For this purpose, CommFind uses a hierarchical agglomerative method, with the additional constraint of merging communities only if they are actually connected in the network.

SpinGlass [48] is another modularity optimization algorithm, which relies on an analogy between community structures and physical spin glass models. Each node is represented by a spin whose state corresponds to the node community index, whereas the network topology is represented by the couplings between spins. The energy level of the model is specified so that the absence of inter-community links and the presence of intra-community links are favored. Under certain conditions, the spin configuration leading to the minimal energy level for this system corresponds to the community structure of maximal modularity. To estimate this ground-state, SpinGlass uses simulated annealing [51], a Monte Carlo optimization method. The global aspect of this method and the non-hierarchical approach of the process are the main differences with the other presented modularity optimization algorithms.

A whole family of algorithms is based on link-centrality measures. The idea behind this approach is that inter-community links are the most central, in the sense one has to use them to go from any node in one community to any node in the other one. On the contrary, given the fact communities are by definition denser

subgraphs, it is likely that a number of paths exist to connect any two nodes located in the same community, making intra-community links less central. In *EdgeBetweenness* [40], this idea is used to implement a hierarchical divisive approach. The process starts with the original network, and iteratively removes the most central links. The network is consequently split into smaller and smaller components, considered as communities in the original network. The process ends when no link remains. Note the centralities are updated at each iteration to take the last removal into account. The algorithm relies on the edgebetweenness centrality, which considers the number of shortest paths going through the link of interest. *Radicci* [47] applies the same process with a different measure called link transitivity. It focuses on triangles rather than shortest paths, considering links belonging to many triangles are more likely to be located inside communities, due to their higher density.

Another approach consists in first defining some function to measure the distance between nodes, and then applying a distance-based clustering approach to estimate communities by minimizing and maximizing intra- and inter-community distances, respectively. *WalkTrap* [49] uses a random walk-based distance, based on the probability to go from one node to the other in a fixed number of steps. Ward's method [32] is then applied to get the communities, which makes WalkTrap a hierarchical agglomerative algorithm. *MarkovCluster* [46] also relies on random walks, but does not include any clustering phase. It iteratively repeats a two-stepped process applied to the network transfer matrix, which contains the probabilities for a random walker to go from one node to another. First, this matrix is raised to some specified power, in order to get a transfer matrix containing probabilities for longer paths. Like for WalkTrap, pairs of nodes with a high probability are supposed to be in the same community. Second, each element in the matrix is raised to some specified power, in order to favor those higher probabilities. The resulting matrix is then normalized to get a new transfer matrix. Both steps are repeated until convergence. The final matrix is binary, and is interpreted as an adjacency matrix describing a network with multiple components. Each one of these components is considered as a community in the original network.

A different approach consists in adopting a data compression perspective and considering the community structure as a set of regularities in the network topology, which can be used to represent the whole network in a more compact way. The best community structure is therefore the one maximizing compactness while minimizing information loss. In *InfoMap* [43], the community structure is represented through a two-level nomenclature based on Huffman coding: one to distinguish communities in the network and the other to distinguish nodes in a community. The problem of finding the best partition is expressed as minimizing the quantity of information needed to represent some random walk in the network using this nomenclature. With a partition containing few inter-community links, the walker will probably stay longer inside communities, therefore only the second level will be needed to describe its path, leading to a compact representation. The authors optimize their criterion using simulated annealing [51].

LabelPropagation [44] relies on a completely different method, based on the simulation of a propagation mechanism. All nodes are initially assigned a different label. Then an iterative process is applied, consisting in giving a node the label which is majority amongst its neighbors (ties are broken randomly), if it is not already the case. The process converges and stops when this condition is verified for all nodes. Communities are then obtained by considering groups of nodes with the same label. By construction, one node has more neighbors in its community than in the others.

Besides the process they implement and the properties of the communities they identify, community detection algorithms also differ in the nature of the information they are able to process. Most of them are limited to unweighted and undirected links. As shown in Table 3, EigenVector can process directed links provided they are unweighted; five algorithms are able to process weighted links provided they are undirected; and only MarkovCluster and InfoMap have the ability to process both weighted and directed links. Gathering the data needed to extract weighted or directed networks is potentially more costly than for plain simple links. We selected algorithms with different abilities, in order to test if such a cost was justified and resulted in substantially different communities on our data.

All the selected implementations are open source and freely available. Table 3 shows which implementations we used: either the program available on the author's website, or the one provided with the R language [31] implementation of the iGraph library [38].

2.2.3 Partition Comparison

To compare the groups estimated by the previously described approaches, we chose the *adjusted Rand index* (ARI), which is widely used to measure similarity between two partitions of a given dataset. The original *Rand index* (RI) [52] is defined as $RI=(a+d)/(a+b+c+d)$, where a (resp. d) corresponds to the number of pairs whose elements belong to the same (resp. different) group(s) in both partitions, and b (resp. c) to the number of pairs whose elements belong to the same group in the first (resp. second) partition, whereas they belong to different groups in the second (resp. first) one. The adjusted version [53] is defined as:

$$ARI=(RI-E)/(1-E) \quad (2)$$

Where E is the amount of similarity expected to be due to chance. The upper limit of this measure is 1 (the two partitions are exactly the same). The value 0 indicates a partial overlap, equivalent to what would be observed if both partitions were random (i.e. $RI=E$). Negative values indicate a strong divergence between the partitions.

The ARI was used to compare the partitions estimated by the various community detection algorithms and reach a reference partition, and to compare the partitions resulting from the cluster analysis with this reference partition.

2.2.4 Predictive Analysis

The second part of our analysis consisted in elaborating a predictive model of the composition of communities, on the basis of the factual attributes. A classic approach for this is to conduct a discriminant analysis, which allows estimating a model taking the form of a set of linear classification functions. Using the community partition as a reference, such a model would allow predicting the community of an instance thanks to its attributes. Its interest is also explicative, since it is possible to analyze the data used for the prediction to characterize the communities. However, this type of analysis was designed to handle numeric data, which is not our case. Extensions such as discriminant correspondence analysis and alternatives such as multinomial sigmoid regression (with Probit or Logit models) allow processing nominal data. But to our knowledge, no tool allows using heterogeneous data such as our factual attributes (some of which are real, ordinal, dichotomous and nominal). This would prevent us from applying the exhaustive exploration of the attributes we planned.

We propose an alternative approach based on the use of the partitions resulting from the cluster analysis. Let us consider a cluster C, and note $u(i)$ the community of one of its instances i according to the community detection result, and $\hat{u}(i)$ the estimation of this community according to our model. By setting $\hat{u}(i)=\text{argmax}_C(u(i))$, we assign all instances in C to the community which is prevalent in this cluster. By applying the same principle to all clusters, we can define a correspondence between each cluster and one of the communities, and consequently predict community membership for all instances.

Note it is possible to have the same community associated to several clusters, which means the clusters are considered as subgroups of this community. It is also possible for a community not to be associated to any clusters, which means its instances were diluted in several clusters having themselves a larger intersection with other communities. With this model, misclassifications appear when an instance belongs to a cluster whose associated community is not the correct one for this instance.

The quality of the model can be assessed by processing its success rate when predicting the communities of all the studied instances. One expects to get a high success rate when the clusters are similar to the communities, or when they form a subdivision of the community partition. Our assumption is the attributes used by the model to successfully predict community membership are characteristic of the considered communities. To explore the attribute space, we will build and evaluate a model for each partition identified during the cluster analysis phase, i.e. we will consider all possible combinations of attributes. The most discriminant one will be identified by selecting the model leading to the highest success rate. These

attributes can then be ranked in terms of explicative power by considering the success rate associated to the models built on any subcombination. For instance, if the best predictions are obtained using three attributes, one can consider the prediction abilities of the models built using one or two of them.

3 Results and Discussion

Our presentation of the results follows the approach we presented in section two. First, we describe and compare the communities identified in our networks. In the following subsection, we explain how they can be used in the context of Business Science. Then, we perform a cluster analysis of the factual data and describe the obtained results. We use certain of these partitions in the following subsection to define our predictive model and identify the most discriminant attributes relatively to the communities. Finally, we use these attributes to interpret the communities.

3.1 Community Detection

Before performing community detection, we cleaned the social network extracted from our relational data by removing its isolated nodes. This is a classic procedure, because all algorithms consider separated components as distinct communities, leading to many meaningless communities when such nodes are present. We additionally removed the very small components to get a single giant component. These operations reduced the number of nodes to 552. We then applied all the appropriate algorithms to four different versions of the network presenting (un)directed and (un)weighted links.

Table 4 shows the modularity values obtained on the unweighted undirected network. Most of the algorithms reach a very high modularity, close to 0.85. We can distinguish Radicchi, EigenVector and MarkovCluster, which are slightly above with a modularity under 0.8. This is still very high if we consider 0.7 as a high value for a real-world network, as previously mentioned.

Among the algorithms able to process weighted undirected networks, FastGreedy, Louvain, SpinGlass, InfoMap and WalkTrap are once again the top ones in terms of modularity. LabelPropagation performance decreases slightly (under 0.8), while MarkovCluster stays at the same level than before, clearly above the others.

On the unweighted directed network, InfoMap does not manage to find any meaningful partition and gets a zero modularity. EigenVector and MarkovCluster have modularities very close to those obtained on the unweighted undirected network. Only two algorithms are able to process the weighted directed networks. MarkovCluster obtains a modularity slightly higher than on the three other networks. Like for the previous network, InfoMap finds only one community and has

a zero modularity. Interestingly, the additional information conveyed by the link direction seems to make it impossible for this algorithm to find a community structure suitable for compression.

Table 4. Quality of the community detection results expressed in terms of modularity

Algorithm	Unweighted		Weighted	
	Undirected	Directed	Undirected	Directed
CommFind	0.8662	-	-	-
EdgeBetweenness	0.8754	-	-	-
EigenVector	0.7962	0.7823	-	-
FastGreedy	0.8780	-	0.8727	-
InfoMap	0.8441	0.0000	0.8384	0.0000
LabelPropagation	0.8165	-	0.7918	-
Louvain	0.8754	-	0.8733	-
MarkovCluster	0.6988	0.6805	0.7053	0.7323
Radicchi	0.7974	-	-	-
SpinGlass	0.8753	-	0.8667	-
WalkTrap	0.8549	-	0.8449	-

To assess the effect of weights and directions, we compared the partitions estimated by each algorithm over the different networks. MarkovCluster, which was applied to all four networks, leads to very similar partitions with ARI values between 0.85 and 0.9. The same remark holds for most of the other algorithms which could be applied to several networks, although with slightly lower ARI values (between 0.78 and 0.9). Obviously, InfoMap is the exception because of its zero modularity for both directed networks. We can conclude from these results that, on our data, considering directions and/or weights during the community detection process does not seem to affect the identified communities. Yet, the selected algorithms are supposed to take advantage of this extra information, so we expected to observe a more significant difference. This might be due to the fact the unweighted undirected network is already highly modular, which means its topology conveys enough information to efficiently discover the community structure: in this case, there is not much to improve by including weights and directions in the process. Another explanation would be the tested algorithms are not able to efficiently take advantage of the extra information.

There is no clear improvement of the performance when considering extra information, so we decided to focus on the unweighted undirected networks, for which we obtained the highest modularity values. We concentrated our analysis on the top four algorithms, which have very close modularity values (above 0.875): EdgeBetweenness, FastGreedy, Louvain and SpinGlass. Interestingly, they lead to partitions with comparable sizes (number of communities): 22 for EdgeBetweenness, FastGreedy and Louvain, 28 for SpinGlass. By comparison, MarkovCluster and EigenVector (modularity bellow 0.8) found 85 and 48 communities, respec-

tively. This is an important point, because for interpretation purpose, it is convenient to have a small number of relatively large communities compared to the size of the network and the number of available attributes.

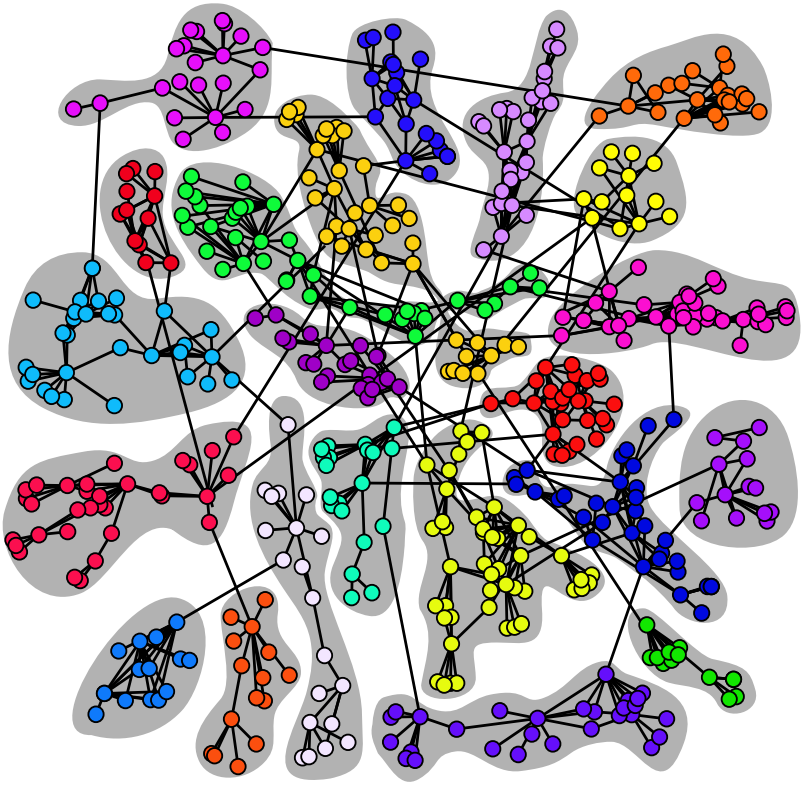

Fig. 1. Communities detected by FastGreedy using the relational data. The modularity processed for this partition is 0.88, i.e. a high value, which has two meanings: the network has a community structure and FastGreedy managed to identify it well. Isolated nodes were discarded for clarity.

We compared the partitions estimated by the top four algorithms using the ARI and obtained values between 0.73 and 0.82, which can be interpreted as a strong agreement between all four algorithms. Thanks to this high similarity between the optimal partitions, we can conclude the detected communities are relatively stable. FastGreedy is particularly interesting, because not only did it lead to the partition with the highest modularity and smallest number of communities, but it also has the highest ARI values when compared to the three other top algorithms. In other terms, it can be considered as a good compromise, which is why we will focus on this *reference partition* in the rest of our analysis. Fig. 1 gives a graphical representation of the trimmed network with the 22 communities identified by FastGreedy.

3.2 Business-Oriented Interpretation

As we mentioned in the introduction, previous works in Business Science have emphasized the role of social networks in information diffusion and purchase decision processes. These can be studied through the analysis of the network structure, and in this context community detection is particularly interesting. But communities are also potentially useful for they are groups of persons. Discovering this kind of groups and characterizing them in terms of purchase (or related) behavior appears as a major Business objective. The presence of such a property would increase even more the informative value of the communities, which is why we examine our results from a behavioral point of view in this section.

Table 5. Characterization of the communities in terms of hobbies and purchase behavior

Community	Hobbies		Mobile Phones		Digital Player
1	Music	Sport	Nokia	-	Apple
2	Cinema	Sport	Nokia	Sony-Ericsson	Samsung
3	-	-	Nokia	-	-
4	Music	-	Nokia	Samsung	Creative
5	Sport	-	Samsung	Nokia	Apple
6	Reading	Music	Samsung	Sony-Ericsson	Apple
7	Cinema	Dance	Samsung	Sony-Ericsson	Apple
8	Sport	-	Samsung	-	Apple
9	-	-	Nokia	-	Apple
10	Cinema	Sport	Nokia	Samsung	Apple
11	Music	-	Nokia	-	-
12	-	-	Sony-Ericsson	-	Apple
13	Music	-	Nokia	Samsung	Sony
14	-	-	Nokia	-	-
15	-	-	-	-	-
16	Photo	-	-	-	-
17	Cinema	Sport	Nokia	Samsung	Philips
18	-	-	Nokia	Sony-Ericsson	-
19	-	-	Nokia	Samsung	-
20	Reading	-	Nokia	Samsung	Apple
21	Theater	-	Nokia	-	-
22	Cinema	Sport	Samsung	Nokia	Apple

We selected three emblematic objects of our student population: its hobbies and two important purchases, i.e. mobile phones and digital players. Our goal was to study the behavior related to these objects and to compare communities through this means. We analyzed the 22 communities to identify the two most widespread hobbies, the two most owned brands of mobile phones, and of digital players (Ta-

ble 5) for each of them. Note it was not possible to find characteristic features for all communities.

The results we have in this first step of our study rely on too little data to be generalized (only 224 respondents). Nevertheless, clear differences between communities, concerning the hobbies and purchased brands, appear overall. There are globally two different kinds of communities for each or the three objects of study. In the first one, there is a unity of tastes and one, or sometimes two objects are clearly preferred. The nature of these objects depends on the considered communities. On the contrary, in the second kind, no clear trend appears, meaning there is no preferred object for this community (which can also be considered as a characteristic in itself).

If we focus on the hobbies, communities 3, 9, 12, 14, 15, 18 and 19 belong to the second kind. Among the remaining communities (first kind), we can regroup a majority of communities sharing the exact same interests, or very close ones. For instance, students in communities 2, 10, 17 and 22 are mainly interested in Cinema and Sport. Of course these are rough categories, and in reality a certain variance can exist: for example people probably do not practice the same sports nor like the same movies. Some communities have uncommon hobbies, for instance Dance and Theater are cited only by communities 7 and 21, respectively.

For the mobile phones, several tendencies can be highlighted. First of all, the brand Nokia is clearly the most widespread brand for the respondents. Nevertheless, important differences exist between Nokia-dominated communities. In communities 1, 9 and 14, nearly all the students own a Nokia mobile phone, whereas others have a second brand which is very often Samsung (4, 10, 13, 17, 19 and 20) and sometimes Sony-Ericsson (2 and 18). The second preferred brand is clearly Samsung, which is even dominant in a few communities (5-8 and 22). Interestingly, Apple does not appear, whereas it is largely dominating digital players. Only two communities do not have dominant mobile phone brand (15 and 16), which emphasizes the importance of brands in this sector, and the relevance of community detection.

Concerning the digital players, the most owned brand is by far Apple, which dominates in most of the communities. Interestingly, among the four communities interested primarily in music, Apple is dominant in only one (1), whereas Creative and Sony are prevalent in one community each (resp. 4 and 13). The remaining one (11) could not be characterized by any brand.

This raw analysis is just a first work on our data from a Business Science perspective. Further analyses need to be performed to uncover more meaningful information. For instance, one could be interested to know how hobbies correlate with the brands of purchased items. It is however perfectly illustrative of one of the interests of charactering communities in such a way: one can specifically select appropriate targets from a marketing perspective. Moreover, the structure of the network allows using more precise approaches, for instance by targeting a few persons depending on their connections.

Communities are valuable for Business Science analysis, but as mentioned before, obtaining them requires relational data, which are costly compared to the in-

dividual ones. For this reason, in the rest of this section we present our results regarding the identification of clusters based on factual individual data, and their usefulness to predict community membership.

3.3 Cluster Analysis

As we explained in the methods section, we decided to perform the cluster analysis in an exploratory way. For this reason, we applied all five selected clustering algorithms to all possible combinations of factual attributes. One could think a way of discriminating the numerous resulting partitions would be to select the few ones with the most separated clusters. However, this is not possible due to the nature of our data, which contains several nominal and dichotomous attributes. For some of these attributes, there was no other possibility to compare them than to define binary dissimilarity functions. Yet, when a clustering algorithm is presented a combination of nominal attributes compared via such a function, it will necessarily lead to a perfect partition, containing as many clusters as there are distinct combinations of values for the considered attributes. For instance, if we consider only the gender and entrance examination (both dichotomous attributes), we will obtain four perfectly separated clusters. Moreover, our goal is to use the clusters to predict community membership. These are the reasons why we selected cluster partitions depending on how much they fit the reference community partition.

Table 6. Three best combinations of attributes in terms of fitting of the cluster partition to the community partition. Values correspond to the fit quality expressed using the ARI.

Attributes	Agnes	DBScan	Diana	Pam	TwoStep
Department, Class, Grade	0.457	0.444	0.450	0.456	0.448
Department, Class, Grade, Highschool	0.433	0.223	0.424	0.422	0.429
Department, Class, Highschool	0.413	0.215	0.413	0.422	0.405

Table 6 shows the results obtained for each algorithm over the best three combinations of attributes. The fit with the reference community partition was measured using the ARI. Other combinations lead to much lower ARI values (bellow 0.32 and mostly close to zero), which is why they are not presented here. For all five algorithms, clusters estimated using only the Department, Class and Grade attributes are the closest to the reference community partition, leading to ARI values close to 0.45. These values are intermediary, which seems to indicate the individual attributes used during the clustering process contain a part of the information underlying the network community structure. So on the one hand, we were able to use individual data to identify clusters which are relatively close to (or rather: not significantly different from) the communities estimated from relational data. But on the other hand, the corresponding partitions also have intermediate quality when considering how well they separate the space of attributes. This

seems to confirm our previous observations regarding the difference in the nature of the information conveyed by the individual and relational data.

The partitions estimated by Agnes, DBScan, Diana, Pam and TwoStep for the first combination of attributes contain 25, 38, 21, 20 and 22 clusters, respectively. Except for DBScan, these sizes are very close to the 22 communities contained in the reference partition. The ARI values processed between the clustering algorithms are extremely high, ranging from 0.92 to 0.98, which means they identified almost the same clusters.

For the two other combinations of attributes presented in Table 6, the overlap with the reference partition is slightly lower. One can notice these combinations differ from the first one only by one attribute, which can be interpreted as a confirmation of the relevance of these attributes to discriminate the communities. The ARI values between the clustering algorithms are still very high, although slightly lower than for the first combination (close to 0.9). This is not true for DBScan however, which fails to detect any relevant clusters.

In summary, the various comparisons we conducted between the groups estimated by the cluster analysis approach and those identified by the community detection algorithms, generally lead to close to zero ARI values, meaning the overlap between the corresponding partitions is very low. However, in some specific cases, appropriate combinations of attributes resulted in ARI values significantly different from 0. Thus, a link, even if a tenuous one, seems to exist between some individual attributes and the communities derived from the relational data. The next step in our study consisted in designing a predictive model to assess in which part the repartition of students in the different communities is determined by the attributes identified in this section.

3.4 Composition of the Communities

The work described in the previous section allowed us to identify the attributes which seemed to be the most relevant to discriminate the communities. In this section, we take advantage of the estimated clusters to design several models able to predict the community of an instance based on certain of its attribute values. Our models are obtained by defining a correspondence between each cluster and a community, as explained in section 2.2.4.

Table 7 presents the success rates for the previously presented cluster partitions, characterized here by the algorithm and combination of attributes used. The best performance is obtained with the DBScan partition on the first combination, with a 81.9% success rate, but the results are also relatively high for the other algorithms (around 72-75%). The ARI values we processed for this partition were clearly higher than zero, but not very high (around 0.45), so we expected success rates much lower than 82%. However, this score has to be dampened. Indeed, with our approach, the number of clusters has an important effect on the prediction success rate. For instance, a partition in which all clusters contain only one instance

each (which is a very bad clustering result) will necessarily lead to a perfect prediction. This explains why DBScan (38 clusters) has a higher success rate than the other algorithms (around 20 clusters) for the first combinations of attributes, when it has a lower ARI value. Moreover, the prediction can be applied only to students for which the considered attributes are available. Yet, the grade was a facultative question, and many students did not give any answer (only 96 responses). So, even if the success rate is high, it only concerns a few cases. Finally, the grade question is a delicate one, in the sense students with bad grades are less likely to answer it, which would bias our results.

Table 7. Success rates obtained by applying our predictive approach to the previously described cluster partitions

Attributes	Agnes	DBScan	Diana	Pam	TwoStep
Department, Class, Grade	75.5%	81.9%	72.3%	72.3%	73.4%
Department, Class, Grade, Highschool	69.7%	47.9%	64.8%	69.0%	68.3%
Department, Class, Highschool	71.8%	46.9%	71.8%	70.4%	71.1%

For both other combinations of interest, the best results are obtained with the Agnes partition: 69.7% and 71.8%, respectively. This means we obtain a success rate of more than 68% on the basis of bigger samples, using combinations which still include Department and Class.

In order to understand how the attributes of interest compare in terms of explicative power, we additionally examined them separately using exactly the same method. The best result is obtained with Class (56.5% for all five algorithms), followed by Grade (43.6% with TwoStep) and Department (26.9% for all five algorithms). The predictive rate of Highschool alone is very low: 9.7% with all five algorithms. These scores allow us to rank these attributes of interest in terms of discriminant power.

3.5 Interpretation of the Discriminant Attributes

We *a priori* supposed the Department would be the most discriminant attribute, for the communities, because students from the same department spend most of their time together, making it easy to develop new relationships and strengthen older ones. However, the Class attribute happens to have the best predictive power, far beyond the department. This could be explained by a specificity of the GSU. Students integrating this university come from all parts of Turkey; they have very different skills and levels, both from the academic and linguistic perspectives. In particular, some of them have been speaking French since nursery school, whereas others never practiced this language before entering the GSU. For this reason, before starting the actual Lisans degree, they must follow a preparatory class for one or two years, including an intensive French course. Most students re-

cruited via the national examination do not speak French and have to follow the two-year-long preparation. All students which succeeded in the internal examination speak French and are prepared for only one year. During these preparatory years, all departments are mixed, because students do not follow specialized classes yet, but only French, methodological and common-core classes. We suppose this mixing make students develop cross-departmental relationships more easily. Moreover, this takes place during the first university years, and in a very new context for many students: they are far from home (Turkey is a large country), and family is very important in the Turkish society. For all these reasons, we think these relationships last even after the end of the preparatory program, when students enter specific departments and are separated from most of their preparatory fellows. Furthermore the university campus is very small, and there is therefore no major spatial obstacle to the persistence of interdepartmental relationships.

The Grade is the second best predictive attribute, which could be interpreted as the fact students with similar notes tend to get closer. However, as we stated before, this has to be interpreted carefully because of the scarcity of the responses concerning the corresponding question. The third discriminant attribute (Department) was also the most predictable, because spending daily hours together, working, interacting and sharing the same classroom make people closer and is favorable to the apparition of strong relationships (be it friendship or enmity). Moreover, students belonging to the same department are supposed to have the same academic interests, which can result in easier bounding and common club activity.

Some factors have surprisingly no influence on the repartition between communities. Gender, which could be supposed to have a central role in student interaction, especially in the case of young people evolving in a new environment, far from their home and family, does not seem to affect the way communities are formed. This is all the more surprising that stereotypes about Turkey often show some conservative part of the population considers gender separation as very important. Regarding the GSU students, this situation is not uncommon at all, although the majority of them come from a rather liberal background.

At a lesser degree, we expected high-school specialization (mathematics, literature…), and home city to have a noticeable effect. The life in Istanbul is clearly more cosmopolite, liberal, and culturally richer than in most parts of Turkey. For this reason, persons living in Istanbul sometimes have some sort of superior attitude towards people coming from rural environments, which can be aggravated by the differences of social status, and we expected this to appear in the composition of communities.

Of course, the scope of all these comments must be qualified, because they hold only for the students for which the mentioned discriminant attributes allow correctly predicting the community. For the rest of the population, we were not able to find a way to match communities and attributes, that is to find a link between the relational and individual data. This goes in the same direction than our results from the group comparisons, i.e. relational and individual data seem to convey different information, at least partially.

4 Conclusion

In this work, we gathered data from a population of university students, using a survey. From these data, we retained only the individual factual and relational information, and a part of the individual behavioral information, leaving the rest of the individual information for further exploration. We extracted a social network from the relational information, and detected 22 communities. Each community is characterized by a denser interconnection compared to the rest of the network. Using the individual behavioral information, we illustrated how these communities can be used to extract meaningful information in the context of Business Science. We characterized the communities in terms of hobbies and purchase behavior for two groups of products (mobile phones and digital players). We reached the conclusion that if most communities have heterogeneous tastes, it is nonetheless not the case for all of them. Detecting these communities and identifying their characteristics is a major asset for Business Science. Moreover, communities are supposed to have a strong effect on the buying process and decision making [12]. Consequently, any mean able to provide more information regarding potential client membership is extremely relevant and worthwhile.

For this purpose, in an attempt to predict communities from individual data, we then focused on the analysis of our individual factual information through classic cluster analysis. We considered all possible combinations of the factual attributes and compared the resulting partitions with the community partition, using the adjusted Rand index (ARI) [53]. This exploratory approach allowed us to identify the combinations of attributes leading to the best cluster partitions, in terms of similarity with the community partition. The corresponding ARI values are close to 0.45, which means although there is clearly more than random overlapping between clusters and communities, there is nevertheless a significant difference between them. Additionally considering these clusters were poorly separated, we concluded that, on these data, the information conveyed by the relational and factual individual data seem to differ significantly, at least in terms of groups of students. In other terms, the analysis traditionally performed in Business Science do not allow accessing the same information than what can be obtained through community detection. To our knowledge, this type of comparison was never performed before, especially in the domain of Business Science.

The third part of our analysis aimed at identifying the most discriminant factual attributes relatively to the communities. Our goal was to try giving an attribute-based interpretation to these groups formed only thanks to topological information. For this purpose, we took advantage of our results from the cluster analysis to build a predictive model, using the communities as reference groups. We found out the year of study (Class), the current grade (Grade) and the current department (Department) were the attributes of interest. We proposed some interpretations regarding why a community can be correctly predicted from these attributes.

5 Future Work

Our main result was to show the importance of community detection in a Business Science context. Indeed, not only are communities central in the information diffusion process, but they can also be characterized in terms of purchasing behavior. Further studies on other data and domains will show if they can be the basis of a new criteria for market segmentation. The perspectives for Business and Marketing Sciences seem very exciting and promising.

However, we consider this work as a first step in the analysis of our field results. Consequently, it suffers from some limitations we plan to solve quickly in the forthcoming articles. First, we mainly focused on the factual part of the individual information. The next step will consist in taking into account the rest of the behavioral and/or the sentimental individual data: maybe they convey the information necessary to define clusters exhibiting a better overlap with the communities. Second, we estimated mutually exclusive communities, which does not seem realistic, in the sense people often belong simultaneously to several groups. This could be solved by using an appropriate community detection algorithm [54] and fuzzy cluster analysis. Third, from a more general perspective, we plan to deepen our understanding of the way communities are constituted by proposing a dynamic model of community building.

Our work also suffers structural weaknesses, inherent to the context of the study. First, the survey was conducted in a small institution, with specific characteristics, which makes it difficult to generalize our results to another situation. This problem could be addressed by performing similar studies in other contexts, and for this purpose we are trying to start collaborations with searchers from other universities. Second, and more importantly, the data we analyzed is far from being complete, since it represents a relatively small part of the total number of students in the GSU. This response rate is normal for such a survey, especially considering the fact students participate on their behalf only. To improve this rate, we conducted our survey again, one semester later. This should both provide additional respondents, and add an interesting dynamic dimension for those who participated to both surveys.

Acknowledgments We would like to thank Günce Orman, who helped organizing and translating the survey, Siegfried Devoldère who also translated parts of the questions, and Taleb Mohamed El Wely who programmed the electronic form and designed the survey website. Our gratitude also goes to the reviewers, who provided us constructive comments and allowed us to improve the quality of this article.

References

1. Baret, C., Huault, I., Picq, T.: Management et réseaux sociaux: Jeux d'ombres et de lumières sur les organisations. Revue Française de Gestion **32**(163), 93-106 (2006).

2. Comet, C.: Productivité et réseaux sociaux: Le cas des entreprises du bâtiment. Revue Française de Gestion **32**(163), 155-169 (2006).
3. Simon, F., Tellier, A.: Créativité et réseaux sociaux dans l'organisation ambidextre. Revue française de gestion **187**, 145- 159 (2008).
4. Ferrary, M.: Apprentissage Collaboratif et réseaux d'investisseurs en capital-risque. Revue Française de Gestion **163**, 171-181 (2006).
5. Ranie-Didice, B.: Capital social des dirigeants et performance des entreprises. Revue des Sciences de Gestion **231/232**, 131-135 (2008).
6. Fondeur, Y., Lhermitte, F.: Réseaux sociaux numériques et marché du travail. Revue de l'Ires **52**(3), 102-131 (2006).
7. Guieu, G., Meschi, P.-X.: Conseils d'administrations et réseaux d'administration en Europe. Revue française de gestion **185**, 21-45 (2008).
8. Dwyer, P.: Measuring the value of electronic word of mouth and its impact in consumer communities. Journal of Interactive Marketing **21**(2), 16 (2007).
9. Goldenberg, J., Han, S., Lehman, D.R., Hong J.W.: The role of hubs in the adoption process. Journal of Marketing **73**, 1-13 (2009).
10. Steyer, A., Garcia-Bardidia, R., Quester, P.: Modélisation de la structure sociale de groupes de discussion sur Internet: implications pour le contrôle du marketing viral. Recherches et Applications en Marketing **22**(3) (2007).
11. van der Merwe, R., van Heerden, G.: Finding and utilising opinion leaders: social networks and the power of relationships. South Africa J. Business Management **40**(3), 65-73 (2009).
12. Hyunsook, K.: Comparing fashion process networks and friendship networks in small groups of adolescents. Journal of Fashion Marketing and Management **12**(4), 545-564 (2008).
13. Hartmann, W.R., Manchanda, P., Nair, H., Hosanagar, K., Tucker, C.: Modeling social interactions : identification, empirical methods and policy implications,. Marketing Letter(19), 287-304 (2008).
14. Watts, D.C., Dodds, P.S.: Influentials, networks and public opinion formation,. Journal of Consumer Research **34**, 441-458 (2007).
15. Iacobucci, D., Hopkins, N.: Modeling dyadic interactions and networks in marketing. Journal of Marketing Research **29**(1), 5-20 (1992).
16. Burt, R.: Structural Holes and Good Ideas. Am. J. Sociol. **110**(2), 349-399 (2004).
17. Perry-Smith, J.E.: Social yet creative: The role of social relationships in facilitating individual creativity. Academy of Management Journal **49**(1), 85-101 (2006).
18. Rose, D., Charbonneau, J., Carrasco, P.: La constitution de liens faibles: une passerelle pour l'adaptation des immigrantes centro-americaines mères de jeunes enfants a Montréal Canadian Ethnic Studies **33**(1), 73-91 (1999).
19. Granovetter, M.: The Impact of Social Structure on Economic Outcomes. Journal of Economic Perspectives **19**(1), 33-50 (2005).
20. Sureh, C., Srividya, G., Swetha, K.: Viral distribution potential based on active node identification for ad distribution in viral networks. Int. J. Mobile Market., 48-56 (2009).
21. Chollet, B.: L'analyse des réseaux personnels dans les organisations : quelles données utiliser ? Revue Finance Contrôle Stratégie **11**(1), 105-130 (2008).
22. Doyle, S.: The role of social networks in marketing. Journal of Database Marketing & Customer Strategy Management **15**, 60-64 (2007).
23. Droulers, O., Roullet, B.: Emergence du neuromarketing: apports et perspectives pour les praticients et les chercheurs. Décisions Marketing(46), 9 - 22 (2007).
24. Ohme, R., Reykowska, D., Wiener, D., Choromanska, A.: Application of frontal EEG asymmetry to advertising research. Journal of Economic Psychology **31**(5), 785 - 794 (2010).
25. Parlebas, P.: Sociométrie, réseaux et communication. PUF, Paris, FR (1992)
26. Evrard, Y., Pras, B., Roux, E.: MARKET: Etudes et recherches en Marketing. (2000)
27. Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles **37**, 547-579 (1901).
28. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, US-NY (1990)

29. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Paper presented at the International Conference on Knowledge Discovery and Data Mining,
30. Zhang, T., Ramakrishnon, R., Livny, M.: BIRCH: An Efficient Data Clustering Method for Very Large Datebases. Paper presented at the International Conference on Management of Data, Montreal, CA,
31. R Development Core Team: R: A Language and Environment for Statistical Computing. In. R Foundation for Statistical Computing, Vienna, Austria, (2009)
32. Ward, J.H.: Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc **58**(301), 236-244 (1963).
33. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E **69**(2), 026113 (2004).
34. Tasgin, M., Herdagdelen, A., Bingol, H.: Community Detection in Complex Networks Using Genetic Algorithms. arXiv **0711.0491** (2007).
35. Newman, M.E.J.: Modularity and community structure in networks. PNAS USA **103**(23), 8577-8582 (2006).
36. Newman, M.E.J.: Analysis of weighted networks. Phys Rev E **70**(5) (2004).
37. Leicht, E.A., Newman, M.E.J.: Community Structure in Directed Networks. Phys Rev Lett **100**(11), 118703 (2008).
38. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal **695**(Complex Systems) (2006).
39. Donetti, L., Munoz, M.A.: Detecting network communities: a new systematic and efficient algorithm. J Stat Mech(10), P10012 (2004).
40. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. PNAS **99**(12), 7821-7826 (2002).
41. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys Rev E **74**(3), 036104 (2006).
42. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Phys Rev E **69**(6), 066133 (2004).
43. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. PNAS **105**(4), 1118 (2008).
44. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E **76**(3), 036106 (2007).
45. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J Stat Mech, P10008 (2008).
46. van Dongen, S.: Graph clustering via a discrete uncoupling process. SIAM J Matrix Anal Appl **30**(1), 121-141 (2008).
47. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. PNAS **101**(9), 2658-2663 (2004).
48. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. Phys. Rev. E **74**(1), 016110 (2006).
49. Pons, P., Latapy, M.: Computing communities in large networks using random walks. Computer and Information Sciences - Iscis 2005, Proceedings **3733**, 284-293 (2005).
50. Barnes, E.R.: An algorithm for partitioning the nodes of a graph. SIAM Journal on Algebraic and Discrete Methods **3**, 541–550 (1982).
51. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by Simulated Annealing. Science **220**(4598), 671-680 (1983).
52. Rand, W.M.: Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association **66**(336), 846-850 (1971).
53. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**(1), 193-218 (1985).
54. Derenyi, I., Palla, G., Vicsek, T.: Clique percolation in random networks. Phys Rev Lett **94**(16) (2005).